Towards Video Captioning with Naming: A Novel Dataset and a Multi-modal Approach

Stefano Pini, Marcella Cornia(✉), Lorenzo Baraldi, and Rita Cucchiara

Dipartimento di Ingegneria "Enzo Ferrari",
Università degli Studi di Modena e Reggio Emilia, Modena, Italy
stefanopi.93@gmail.com,
{marcella.cornia,lorenzo.baraldi,rita.cucchiara}@unimore.it

Abstract. Current approaches for movie description lack the ability to name characters with their proper names, and can only indicate people with a generic "someone" tag. In this paper we present two contributions towards the development of video description architectures with naming capabilities: firstly, we collect and release an extension of the popular Montreal Video Annotation Dataset in which the visual appearance of each character is linked both through time and to textual mentions in captions. We annotate, in a semi-automatic manner, a total of 53k face tracks and 29k textual mentions on 92 movies. Moreover, to underline and quantify the challenges of the task of generating captions with names, we present different multi-modal approaches to solve the problem on already generated captions.

Keywords: Video captioning · Naming · Datasets · Deep learning

1 Introduction

Video Captioning is a fundamental achievement towards machine intelligence, as it links together vision and language. The task has been gaining a lot of attention in the past few years, thanks to the spread of Deep Learning approaches [10,24] and large scale movie description datasets [15,21]. While current state-of-the-art approaches seem to have brought this task in a rather mature stage, and the computer vision community is focusing on novel techniques for including semantics [14] and enhance the quality of descriptions [18], one fundamental aspect is still missing: that of naming characters in captions with their proper names. Indeed, despite the availability of movie description datasets in which captions come with character names, it is usual practice to replace them with a "someone" tag, thus ignoring the naming task during the generation of the caption.

This choice has been well motivated by the need of training video captioning architectures which are not endowed with naming capabilities, in which the presence of proper names would have been more detrimental than useless. Each proper name, indeed, would have become a new entry in the dictionary, without being the network capable of identifying the relationship between the name and

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 384–395, 2017.
https://doi.org/10.1007/978-3-319-68548-9_36

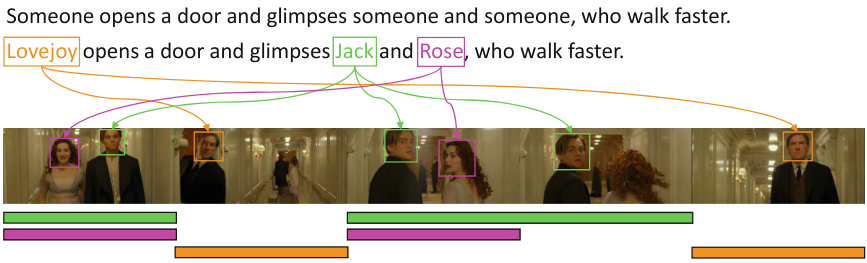

Fig. 1. Generating video captions with character names requires additional supervision besides the sole caption. To this end we augment the M-VAD dataset for movie description by annotating face tracks of movie characters and by associating them with their textual mentions.

the visual appearance of a person, thus making the presence of proper names a driver of noise rather than a useful information. Developing a video captioning network capable of generating captions with proper names in the correct place requires to address several tasks in a deep learning perspective, ranging from face detection and tracking, to face recognition with respect to a given set of names. Moreover, the generative recurrent architecture of such a model needs to be aware of the linguistic structure of the caption being generated, in order to identify the need of outputting a proper name, and then selecting it from a list of possible candidates by exploiting face identification features.

Training a video captioning architecture with naming capabilities by relying on the caption as the sole source of supervision would be particularly challenging, due to the lack of supervision between the textual and the visual domain. In each video, indeed, different faces can appear, only some of them mentioned in the ground-truth caption, and possibly some not even appearing in the cast of the movie, such in the case of background actors. A more grounded form of annotation is therefore needed, to link the visual appearance of a face to the list of characters playing the movie, and to link the mention of a character in a caption with the visual appearance of the corresponding face (Fig. 1).

In this paper, we present two contributions towards the development of video description architectures with naming capabilities: firstly, we augment one of the most popular movie description datasets, the Montreal Video Annotation dataset [21], with the annotations needed to train a neural architecture with naming capabilities. Each video of the dataset has been endowed with face tracks annotations, manually associated with characters of the movie and with textual mentions, thus closing the gap between the textual and the visual domain. Also, we release appropriate training, validation and test splits, which are carefully built by taking into account the specific challenges of the task. To our knowledge, this is the only publicly available dataset providing annotations for video captioning with naming. Secondly, we present a principled use case of the dataset by developing different multi-modal approaches which can solve the naming problem from already generated captions. Experimental results will enlighten and quantify many of the challenges associated with the task, and devise appropriate solutions.

2 Related Work

Our work is related to the generation of captions for video, and to the task of linking visual tracks to names.

The generation of natural language descriptions of visual content has received large interest since the emergence of recurrent networks, either for single images [7], user-generated videos [3] or movie clips [14,24]. First approaches described the input video through mean-pooled CNN features [25], or sequentially encoded by a recurrent layer [3,24]. Other works have then followed this kind of approach, either by incorporating attentive mechanisms [26] in the sentence decoder, by building a common visual-semantic emebedding [10], or by adding external knowledge with language models [23] or visual classifiers [14].

On a different note, the problem of linking people with their names has been tackled in different previous works [2,4,13,19,20], which rely on alignment of video to TV scripts. The goal is to track faces in the video and assign names to them. For example, [4,19] rely on the availability of aligned subtitles and script texts to associate TV or movie characters with their names. In [20] character appearances are modelled as a Markov Random Field, integrating face recognition, clothing appearance, speaker recognition and contextual constraints in a probabilistic manner. [2] propose a discriminative weakly supervised model jointly representing actions and actors in video. [13] tackle the problem of naming people in a video by including ambiguous mentions of people such as pronouns and nominals (i.e. co-reference resolution). [11] present a multiple instance learning based approach which focuses on recognizing background characters showing significant improvement over prior work.

Finally, in [16], authors address the problem of generating video descriptions with grounded and co-referenced people, by proposing a novel dataset and a deeply-learned model. This task significantly differs from the one tackled in this paper, as it aims at predicting the spatial location in which a given character appears, and at producing captions with proper names in the correct place.

3 The M-VAD Names Dataset

Providing the annotations needed to train novel movie description architectures capable of naming characters requires to annotate the visual appearance of each character of a movie, to track it through time, and to annotate the association between visual appearances, character names and textual mentions in the corresponding captions.

To this end, we collect new annotations for the Montreal Movie Description dataset [21]. The dataset consists of 84.6 h of video from 92 Hollywood movies, for a total of 46,523 video clips with corresponding captions, from which we retain only those that contain at least one character mention. For each movie, we provide manually annotated visual appearances of characters, their face and body tracks, and associations with textual mentions in the captions. The dataset,

which we name *M-VAD Names*[1], is annotated in a semi-automatic way, by leveraging face recognition and clustering techniques, and then manually refined.

Face and Body Tracks Generation. To generate face and body tracks, we first detect faces on video frames by using the cascaded approach of [27], in which faces are detected by means of a multi-task Convolutional Network which jointly detects facial landmarks and predicts the face bounding box. After a face is detected, we track it in the following frames with the online tracker presented in [1], which learns an adaptive appearance model using Multiple Instance Learning. To account for tracking errors and camera shot changes, we also compute an appearance similarity measure between predictions in two consecutive frames, and we manually stop the tracking when the similarity is under a threshold. In practice, we found that a pixel-wise difference between consecutive detections is robust enough to discard the majority of errors.

For every face found and tracked for 16 consecutive frames, a new track is saved. This choice is also compatible with the temporal extent used in state-of-the-art action descriptors [22]. For every track, we store the face bounding box as well as the body bounding box, which can be used to compute action descriptors. This extended bounding box is generated with a linear transformation of the face bounding box, expanding it to obtain a square with a side of three times the height of the original bounding box with the face centered in the upper portion of the square. This approach is also similar to what has been done in [2].

Table 1. Overall statistics on the M-VAD Names dataset. Numbers of video clips in each split, and number of mentions, appearances and tracks.

	Number	Avg. per movie	Avg. per character
Train videos	17170	187	-
Test videos	2581	28	-
Validation videos	2708	29	-
Mentioned characters	1392	15	-
Annotated characters	908	10	-
Mentions	29862	325	22
Appearances	20631	224	24
Tracks	53665	583	62

Movie Character Annotations. Having detected face and body tracks, we associate them to characters from each movie with a coarse semi-automatic annotation strategy, which is then manually refined. Specifically, we employ a face

[1] The dataset is publicly available at: http://imagelab.ing.unimore.it/mvad-names.

recognition approach inspired to FaceNet [17] and trained on the MS-Celeb-1M dataset [6]. Each face track is projected into a 128-dimensional embedding space which is trained with the objective of giving very similar descriptors to similar faces (i.e. faces of the same character) and diverse descriptors to different faces (i.e. faces of different characters). At this point, an agglomerative clustering is used to cluster the face representations calculated in the previous phase for every movie. In this way, for every movie we obtain N clusters of similar faces associated to the relative video clips and tracks using an unsupervised method. Every cluster is then manually classified either as a character of the movie, as an unknown character (if it is not part of the cast, or it can not be not recognized by the human annotator), or as wrong (if the track does not contain a face due to errors in the face recognition and tracking phase). Every element of the cluster is manually checked and, if needed, moved to the appropriate cluster. At the end of this process, we obtain the annotation of the main characters of every movie and for every sequence of the dataset. This annotation is then used to associate tracks and appearances with textual mentions in the captions. To avoid human error, each annotation is checked by at least two different people.

Training, Validation and Test Splits. The authors of the M-VAD dataset provided official training/validation/test sets by splitting the original set of movies into three disjoint parts, so that captioning algorithms could be trained on a set of movies, and validated or tested on other movies, with only partially overlapped vocabularies. In the case of naming, it is instead important to have clips from the same movies in all splits, so that the visual appearance of a specific character can be learned and applied at testing time. For this reason, we propose and release an official split for the M-VAD Names dataset.

The split has been done randomly, but applying a series of soft constraints. We forced every movie to have 80% of the video clips into the training set, 10% into the validation set and 10% into the test set. Also, we applied the same splitting criterion to characters, to video clips with only one mention, and to video clips with two or more mentions. Ideally, these constraints ensure that the training set contains the 80% of the video clips of each movie, the same percentage of the characters, of clips with one mention, and of clips with two or more mentions. Finally, we gave priority to have at least a video clip for every character in every group set. We have considered the last constraint as the most important one to try to have ax example for every character in every set.

Statistics. Table 1 reports the overall statistics of the collected dataset. As it can be seen, the dataset contains a total of 22,459 video clips, which are divided into appropriate train, test and validation splits. The number of unique characters found in the screenplays is 1,392, while the overall number of mentions is 29,862. We found some errors in the captions of the M-VAD dataset for many films, so the correct values of unique characters and mentions would be lower. Finally, the unique annotated characters are 908. They appear a total of 20,631 times in the video clips and in 53,665 extracted tracks. Average per movie and

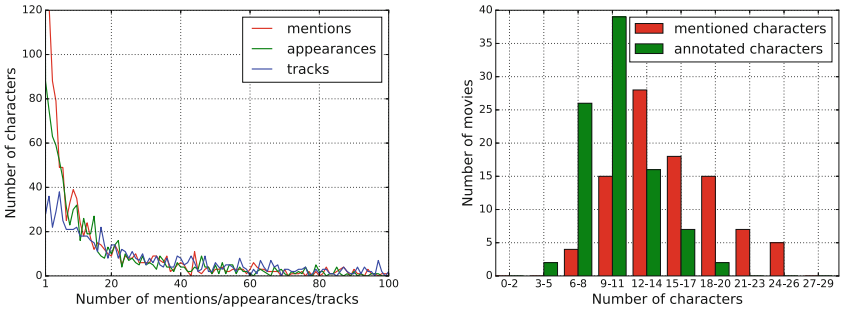

Fig. 2. Statistics on the dataset. Left: distribution of characters with respect to the number of mentions, appearances and tracks. Right: distribution of movies with respect to the number of characters mentioned in the captions and visually annotated.

per character values are reported, to give an estimate of how much data can be extracted from a single film or a single character. Figure 2 reports the distribution of characters with respect to the number of mentions, appearances and tracks on the left, and the distribution of movies with respect to the number of characters mentioned in the captions and visually annotated on the right. Due to errors in the M-VAD mentions and misses during the annotation pipeline, most of the films have less annotated characters than the mentioned ones, but the relative distribution is similar. For the same reasons, there is a peak in the graph on the left for the number of characters with one mention. As it can be observed, lots of characters have less than 20 mentions, annotation and tracks, and only few characters have more than 80.

4 Replacing the "Someone": An Approach to Video Description with Naming

The M-VAD Names dataset provides sufficient annotations to train generative architectures with naming capabilities, which for sure will need to solve the problem of associating textual mentions with character names at generation time. In this Section, we investigate a strictly related task which shares many of the challenges behind that of generating video description with proper names. Indeed, instead of proposing an architecture able to generate captions with names, we tackle the problem of replacing the "someone" tag in existing descriptions with the correct character name.

The task requires to identify for each video clip which character performs the action described in each textual mention in the relative caption. Therefore, it is not only necessary to identify characters present in the scene, but also to understand what each of them is doing. We start from ground-truth descriptions, where we replace each proper name with a "someone" tag, and extract the verb associated with each "someone" subject tag by using a NLP parser [8]. For each video clip, we also exploit face and body tracks from the M-VAD Names dataset,

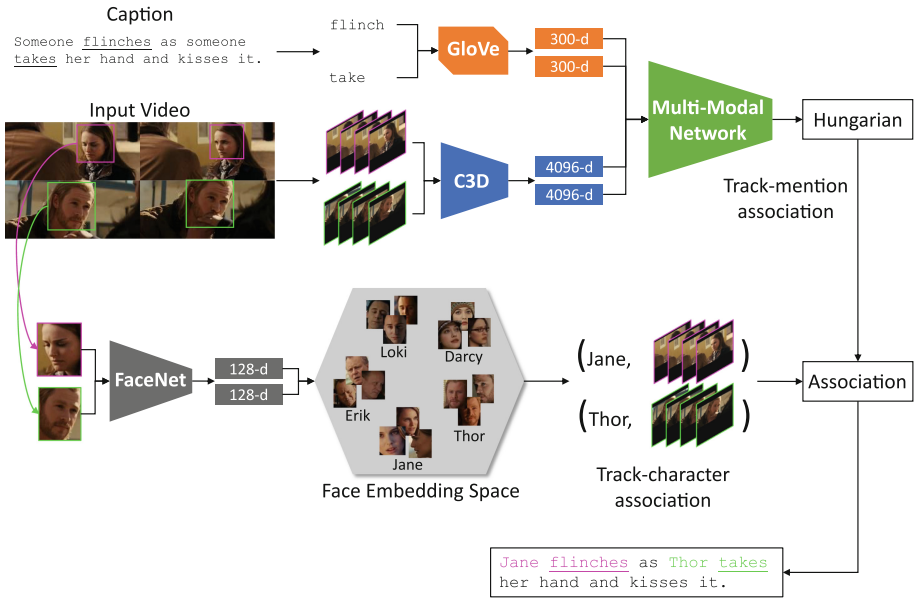

Fig. 3. Summary of our approach for replacing "someone" tags with the correct character names. A Multi-Modal network is trained to predict a matching between visual tracks and actions (expressed by the verbs found in the textual description), while a face recognition approach is used to complete the matching between the track and the character.

and the association with textual mentions. Starting from these information, we build a Multi-Modal network capable of associating a character track with the corresponding mention in the caption, by predicting the similarity between the performed action (i.e. the verb), and the visual track. A summary of the approach is shown in Fig. 3.

4.1 Multi-Modal Network

Given an input track and a verb, we preprocess them to extract appropriate feature vectors, and then feed them to a trained Multi-Modal network which can project visual and textual elements in a common multi-modal space. For tracks, which are 16 frames long, we resize them to 112×112 and extract a 4096-dimensional feature vector from the last but one fully convolutional layer of the C3D network [22], trained on the Sports-1M dataset [9], which encodes motion features computed around the middle frame of the input track. For each verb, instead, we extract a 300-dimensional feature vector by using the GloVe embedding [12] that provides a semantic vector representation for words.

The Multi-Modal network is composed of two different branches for each of the modalities. A visual branch $\phi_v(\cdot)$ takes the track descriptor as input and projects it into a multi-modal space, while a textual branch $\phi_t(\cdot)$ does the same

for the textual counterpart. Training is performed by forcing the network to produce an embedding space such that a track and its corresponding verb have close projections, while a track and a verb which do not match should be far in the embedding space. We do this by introducing two distance terms, $p(\cdot)$ and $n(\cdot)$ to force projections of a verb and track to be close or far by at least a margin M:

$$p(\mathbf{t}, \mathbf{v}) = \|\phi_v(\mathbf{t}) - \phi_t(\mathbf{v})\|_2^2, \qquad n(\mathbf{t}, \mathbf{v}) = \max(M - \|\phi_v(\mathbf{t}) - \phi_t(\mathbf{v})\|_2^2, 0) \quad (1)$$

where $\mathbf{t}$ and $\mathbf{v}$ are respectively a visual track and a verb. In our case, we want each track to be classified as similar to its corresponding verb and dissimilar to a randomly chosen different verb. Additionally, we want to exploit the fact that the same verb could be associated to other tracks and to enforce tracks annotated as "wrong" (i.e. not containing a person) to be classified as dissimilar to that verb. Our final loss function is therefore:

$$L = \sum_{i=1}^{N} p(\mathbf{t}_i, \mathbf{v}_i) + n(\mathbf{t}_i, \mathbf{v}_i^-) + p(\mathbf{t}_i^+, \mathbf{v}_i) + n(\mathbf{t}_i^w, \mathbf{v}_i) \quad (2)$$

where, for each mention in the dataset, $\mathbf{t}_i$ represents the visual track associated with the mention; $\mathbf{v}_i$ the corresponding verb; $\mathbf{v}_i^-$ a randomly chosen verb from the dataset, different from $\mathbf{v}_i$; $\mathbf{t}_i^+$ a second visual track, associated with the same verb $\mathbf{t}_i$ (if present); $\mathbf{t}_i^w$ a "wrong" track.

At test time, we compute the distance between a track and a verb projected into the embedding space (i.e. $p(\mathbf{t}, \mathbf{v})$). Given a video clip, and all the distances between its verbs and its tracks, we compute the optimal assignment between tracks and verbs by using the Kuhn-Munkres algorithm.

4.2 Face Recognition

Once tracks have been assigned to textual mentions, tracks needs to be assigned to characters to complete the task of replacing the "someone". To this aim, similarly to what we did in Sect. 3, we use the 128-dimensional face representation provided FaceNet [17]. Then, we use these representations of the training faces of the characters as the training data of a K-Nearest Neighbours classifier with k equal to 5. In particular, an optimized version of it (kd-tree) is used. This classifier has been chosen to take advantage of the previously generated embeddings and to be sufficiently flexible in the case of characters whose aspect changes during the movie. Clustering classifiers and linear methods, in fact, are not always suitable to classify classes that can have multiple agglomerations in different areas of the space.

5 Experimental Evaluation

We evaluate the performance of the proposed pipeline with respect to different training strategies for the Multi-Modal network, and different classification

Table 2. Experimental results on the "replacing the someone" task, with different loss functions and face recognition approaches. Results are reported on the validation and test split of the M-VAD Names dataset, in terms of accuracy.

	Val. accuracy	Test accuracy
Random assignment	0.111	0.103
MMN Binary with two terms + Face AdaBoost	0.551	0.578
MMN Siamese + Face AdaBoost	0.569	0.593
MMN Triplet + Face AdaBoost	0.529	0.568
MMN Binary with two terms + Face SVM	0.667	0.670
MMN Siamese + Face SVM	0.683	0.685
MMN Triplet + Face SVM	0.630	0.650
MMN Binary with two terms + Face KNN	0.678	0.681
MMN Siamese + Face KNN	0.691	0.701
MMN Triplet + Face KNN	0.648	0.664
MMN Binary with four terms + Face KNN	0.710	0.691
Proposed MMN + Face KNN	**0.715**	**0.712**

approaches for the face recognition part. In particular, besides the proposed loss, we test a Siamese and a Triplet loss function, plus two Binary loss functions. The Siamese loss exploits only the first two terms of Eq. 2, while the Triplet loss is defined as follows:

$$L = \sum_{i=1}^{N} \max(\|\phi_v(\mathbf{t}_i) - \phi_t(\mathbf{v}_i)\|_2^2 - \|\phi_v(\mathbf{t}_i) - \phi_t(\mathbf{v}_i^-)\|_2^2 + M, 0). \tag{3}$$

For the Binary loss functions, instead, we replace $p(\cdot)$ and $n(\cdot)$ with two binary cross-entropy terms, i.e. $p(\mathbf{t}, \mathbf{v}) = -\log d(\phi_v(\mathbf{t}), \phi_t(\mathbf{v}))$ and $n(\mathbf{t}, \mathbf{v}) = -\log(1 - d(\phi_v(\mathbf{t}), \phi_t(\mathbf{v})))$, where d is learnable classification function. In one version, we maintain all the four terms of Eq. 2, and in a second version, we keep only the first two terms, using the re-defined $p(\cdot)$ and $n(\cdot)$.

As mentioned, the architecture of our Multi-Modal network is composed by two different branches: the first one processes the track feature vectors coming from the C3D network, while the second processes the vectorized representations of verbs. In details, the track branch is composed by 4 fully connected layers with 1024, 256, 64 and 16 units respectively, while the verb branch is composed by 3 fully connected layers with 256, 64 and 16 units respectively, all with ReLU activations. Finally, in the case of the Binary loss functions, the outputs of these two branches are concatenated and fed into a fully connected layer with 1 unit and a sigmoid activation, which acts as the learnable classifier d. Weights of all layers are initialized according to [5].

During the training phase, we randomly sample a minibatch containing 32 training samples and we encourage the network to minimize one of the mentioned loss functions through the Stochastic Gradient Descent optimizer. SGD is applied

with Nesterov momentum 0.9, weight decay 0.0005 and learning rate 10^{-2}. The margin M, after cross-validation, has been set to 1 for the Siamese loss, and to 0.2 for the Triplet loss.

5.1 Experimental Results

Experimental results are reported in Table 2. We show the results in term of the final association accuracy between textual mentions and character names, by comparing the proposed Multi-Modal network, described in Sect. 4.1, with the Siamese, Triplet and Binary alternatives. As it can be noticed, the proposed Multi-Modal network obtains the best results on both validation and test sets. Moreover, the performances obtained by the proposed loss and by the Binary loss with four terms confirm the advantage of exploiting wrong tracks and tracks associated with the same verb. For reference, we also show the result of a random replacement of all "someone" tags with a movie character randomly extracted from the character list of each movie.

Moreover, we report the accuracy results by using different face recognition approaches. In particular, we compare the K-Nearest Neighbours classifier (KNN) with the SVM and the Adaboost (with 30 Decision Trees) classifiers. As it can be observed, the KNN performs better than others on both validation and test sets.

6 Conclusion

In this paper we presented two contributions towards the development of movie description architectures capable of indicating characters with their proper names. First, we presented and released an extension of the M-VAD dataset, in which character faces and bodies are manually annotated and associated to textual mentions, thus closing the gap in supervision between the visual and the textual domain. Then, we proposed a multi-modal approach to tackle the task of naming characters on existing descriptions: our approach combines a Deep Network which jointly classifies visual tracks and textual actions, and state-of-the-art face recognition techniques. Experimental results enlighten and quantify the challenges associated with the task of developing novel video description architectures with naming capabilities.

References

1. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR (2009)
2. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: ICCV (2013)
3. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)

4. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is.. Buffy"-Automatic Naming of Characters in TV Video. In: British Machine Vision Conference (2006)
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics (2010)
6. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). doi:10.1007/978-3-319-46487-9_6
7. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: describing novel object categories without paired training data. In: CVPR (2016)
8. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Conference on Empirical Methods in Natural Language Processing (2015)
9. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
10. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: CVPR (2016)
11. Parkhi, O.M., Rahtu, E., Zisserman, A.: Its in the bag: stronger supervision for automated face labelling. In: ICCV Workshops (2015)
12. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (2014)
13. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with "their" names using coreference resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 95–110. Springer, Cham (2014). doi:10.1007/978-3-319-10590-1_7
14. Rohrbach, A., Rohrbach, M., Schiele, B.: The long-short story of movie description. In: Gall, J., Gehler, P., Leibe, B. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 209–221. Springer, Cham (2015). doi:10.1007/978-3-319-24947-6_17
15. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: CVPR (2015)
16. Rohrbach, A., Rohrbach, M., Tang, S., Oh, S.J., Schiele, B.: Generating descriptions with grounded and co-referenced people. In: CVPR (2017)
17. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: CVPR (2015)
18. Shetty, R., Rohrbach, M., Hendricks, L.A., Fritz, M., Schiele, B.: Speaking the same language: matching machine to human captions by adversarial training. arXiv preprint arXiv:1703.10476 (2017)
19. Sivic, J., Everingham, M., Zisserman, A.: Who are you?-learning person specific classifiers from video. In: CVPR (2009)
20. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: Knock! Knock! Who is it? Probabilistic person identification in TV-series. In: CVPR (2012)
21. Torabi, A., Pal, C., Larochelle, H., Courville, A.: Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070 (2015)
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV (2015)

23. Venugopalan, S., Hendricks, L.A., Mooney, R., Saenko, K.: Improving LSTM-based video description with linguistic knowledge mined from text. In: Conference on Empirical Methods in Natural Language Processing (2016)
24. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: ICCV (2015)
25. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: North American Chapter of the Association for Computational Linguistics (2014)
26. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: ICCV (2015)
27. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascaded convolutional networks. arXiv preprint arXiv:1604.02878 (2016)